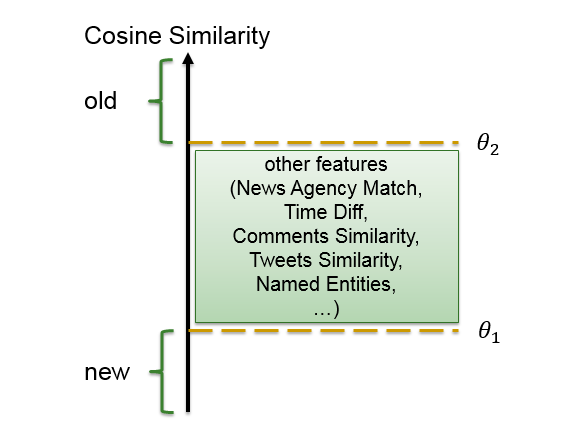
بسم‌الله الرحمن الرحیم

دو هفته اخیر در مورد تأثیر روش وزن دهی روی فرآیند کشف زیر رخداد جدید، فعالیت‌هایی انجام شد که در ادامه خلاصه‌ای از این فعالیت‌ها شرح داده می‌شود:

* برای پیدا کردن دو آستانه‌ای که جلسه قبل در مورد آن صبحت شد، شکل زیر را در نظر بگیرید:



برای تعیین این دو آستانه، از بیشینه و کمینه شباهت کسینوسی هر سند با تمام اسناد قبلی آن استفاده‌شده است: بیشینه شباهت کسینوسی با روش وزن دهی TF.IDF در تمام اسنادی که برچسب old دارند به‌عنوان آستانه شماره 2 و نیز، کمینه شباهت کسینوسی تمام اسنادی که برچسب new دارند، به‌عنوان آستانه شماره 1 در نظر گرفته شد.

* به‌منظور بررسی خطاهای افزوده‌شده(سندهایی که قبلاً درست دسته‌بندی‌شده و الان اشتباه برچسب می‌خورند) از 2 روش انتخاب آستانه استفاده‌شده است.

نتایج به شرح زیر است:

|  |  |
| --- | --- |
| TP | 4 |
| TN | 44 |
| FN | 1 |
| FP | 0 |
| no-decision | 55 |
| Total | 104 |

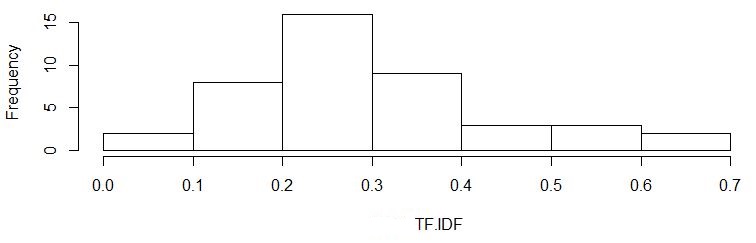
شکل 1 نتایج روش TF.IDF با آستانه اولیه

تعداد کل اسناد 104 سند است که در این میان، 1 سند به‌اشتباه "تکراری" تشخیص داده‌شده، و 55 سند در این روش باید توسط فاز دوم، وزن دهی شوند.

اگر آستانه‌ها را اندکی تغییر دهیم و آسان‌گیرانه‌تر فرض کنیم به نتایج زیر می‌رسیم:

آستانه‌های جدید به شرح زیر است:

دلیل اینکه تغییرات بیشتری نسبت به دارد را می‌توان به توزیع شباهت اسناد ارتباط داد:

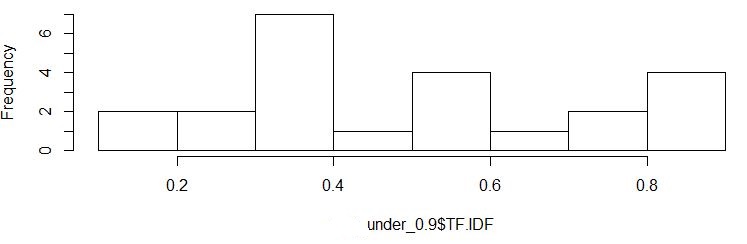


شکل 2 هیستوگرام شباهت کسینوسی اسنادی که برچسب new دارند

در شکل 2، محور افقی، نمایانگر شباهت کسینوسی هر سند با شبیه‌ترین سند قبلی خود در میان تمام اسناد قبلی است.

اولیه، در این مجموعه داده، برابر 0.16 در نظر گرفته‌شده است. همان‌طور که از شکل 2 می‌توان استنباط کرد، با دو برابر کردن این آستانه، می‌توان تعداد بیشتری از اسناد new را شناسایی کرد.

اما در مورد با در نظر گرفتن توزیع، می‌توان به این نتیجه رسید که نمی‌توان با تغیر زیاد آستانه، بهبودی حاصل کرد:



شکل 3 هیستوگرام شباهت کسینوسی اسنادی که برچسب old دارند

با این تفاسیر، نتایج TF.IDF با آستانه‌های جدید به شرح زیر است:

|  |  |
| --- | --- |
| TP | 27 |
| TN | 47 |
| FN | 3 |
| FP | 5 |
| no-decision | 22 |
| Total | 104 |

شکل 4 نتایج روش TF.IDF با آستانه تغییریافته

به‌منظور مقایسه، نتایج دو روش در شکل زیر آمده است:

|  |  |
| --- | --- |
| TF.IDF و آستانه اولیه | TF.IDFو آستانه تغییریافته |
| |  |  | | --- | --- | | TP | 4 | | TN | 44 | | FN | 1 | | FP | 0 | | noDecision | 55 | | Total | 104 | | |  |  | | --- | --- | | TP | 27 | | TN | 47 | | FN | 3 | | FP | 5 | | noDecision | 22 | | Total | 104 | |

خوشبختانه تعداد اسناد باقیمانده که در فاز اولیه برچسب زده نشده‌اند دارای bias معناداری به سمت برچسب خاصی نیستند.

در فاز بعدی، اگر ویژگی‌های فاصله زمانی و یکسان بودن خبرگزاری را در نظر بگیریم،

8 سند برچسب می‌خورند که از این تعداد، 7 سند برچسب درست می‌گیرند و یک سند برچسب اشتباه می‌گیرد که این برچسب اشتباه به FP اضافه می‌شود(معادل False Alarm که وزن آن 0.1 وزن هر FN است)، نتایج جدید به شرح زیر است:

|  |  |
| --- | --- |
| TP | 34 |
| TN | 47 |
| FN | 3 |
| FP | 6 |
| noDecision | 14 |
| Total | 104 |

شکل 5 نتایج روش TF.IDF با آستانه تغییریافته و با استفاده از ویژگی‌های شباهت خبرگزاری و فاصله زمانی

تطابق نظیر به نظیر اسناد با روش پایه:

|  |  |
| --- | --- |
| **Baseline** | **Last Try** |
| miss in DocID :8 | No Error |
| miss in DocID :9 | No Error |
| miss in DocID :11 | No Error |
| miss in DocID :14 | No Error |
| miss in DocID :22 | No Error |
| miss in DocID :28 | No Error |
| miss in DocID :30 | No Error |
| miss in DocID :33 | miss in DocID :33 |
| miss in DocID :35 | No Error |
| miss in DocID :51 | No Error |
| miss in DocID :55 | No Decision |
| false in DocID :71 | false in DocID :71 |
| No Error | false in DocID :89 |
| miss in DocID :106 | No Error |
| miss in DocID :131 | miss in DocID :131 |
| false in DocID :135 | false in DocID :135 |
| miss in DocID :140 | No Error |
| miss in DocID :154 | No Error |
| miss in DocID :165 | No Error |
| miss in DocID :177 | miss in DocID :177 |
| No Error | false in DocId :178 |
| No Error | false in DocID:179 |
| miss in DocID :183 | No Error |
| miss in DocID :194 | No Error |

معنی رنگ‌ها در جدول فوق:

|  |  |
| --- | --- |
| بهبود |  |
| نتیجه یکسان |  |
| افزودن خطا |  |